



# Semi Automatic Construction of ShEx and SHACL Schemas

Iovka Boneva, Jérémie Dusart, Daniel Fernández Alvarez, Jose Emilio Labra Gayo

## ► To cite this version:

Iovka Boneva, Jérémie Dusart, Daniel Fernández Alvarez, Jose Emilio Labra Gayo. Semi Automatic Construction of ShEx and SHACL Schemas. 2019. hal-02193275

**HAL Id: hal-02193275**

**<https://hal.science/hal-02193275>**

Preprint submitted on 24 Jul 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Semi Automatic Construction of ShEx and SHACL Schemas

Iovka Boneva<sup>1</sup>, Jérémie Dusart<sup>2</sup>, Daniel Fernández Álvarez, and Jose Emilio Labra Gayo<sup>3</sup>

<sup>1</sup> University of Lille, France

<sup>2</sup> Inria, France

<sup>3</sup> University of Oviedo, Spain

**Abstract.** We present a method for the construction of SHACL or ShEx constraints for an existing RDF dataset. It has two components that are used conjointly: an algorithm for automatic schema construction, and an interactive workflow for editing the schema. The schema construction algorithm takes as input sets of sample nodes and constructs a shape constraint for every sample set. It can be parametrized by a schema pattern that defines structural requirements for the schema to be constructed. Schema patterns are also used to feed the algorithm with relevant information about the dataset coming from a domain expert or from some ontology. The interactive workflow provides useful information about the dataset, shows validation results w.r.t. the schema under construction, and offers schema editing operations that combined with the schema construction algorithm allow to build a complex ShEx or SHACL schema.

**Keywords:** RDF validation, SHACL, ShEx, RDF data quality, LOD usability, Wikidata

## 1 Introduction

With the growth in volume and heterogeneity of the data available in the Linked Open Data Cloud, the interest in developing unsupervised or semi-supervised methods to make this information more accessible and easily exploitable keeps growing as well. For instance, *RDF graph summarization* [1] is an important research trend that aims at extracting compact and meaningful information from an RDF graph with a specific purpose in mind such as indexing, obtaining high level descriptions of a given dataset, graph visualization, query optimization, or extracting a schema or an ontology.

The Shapes Constraint Language (SHACL) [6] is the W3C standard for defining constraints for RDF graphs. The Shape Expressions Schemas (ShEx) language [14] has a similar purpose but is based on a slightly different set of operators. Both proposals allow to define a set of named shapes that constrain the RDF nodes and describe their close neighbourhoods. A set of such constraints is called a schema and provides high level information about the structure and contents of an RDF dataset. Such schema is usually constructed with a particular

application in mind. SHACL and ShEx schemas allow to increase RDF datasets usability, as they facilitate RDF validation, understanding and querying.

More recently, several authors have used techniques close to graph summarization in order to automatically produce a SHACL [12,15,10] or a ShEx [2,16,7,4] schema. Nevertheless, shapes generated using noisy or incomplete data may contain incorrect data or imprecisions. Most of the times, the shapes should be reviewed and edited by domain experts in order to ensure they fit with the application requirements. The usability of tools letting a proper edition of those schemas has been identified as a challenge [8]. The relative youth of consolidated constraint languages for RDF data causes that this kind of tools has not been thoroughly investigated yet, but a list of desirable features has been proposed [3].

*Contributions* In the current paper we describe the algorithms and the software components that are necessary for the development of an integrated, highly parametrizable and well automated tool that would allow to users with different levels of expertise to construct SHACL and ShEx schemas from data. As a first contribution, we define an algorithm that takes as input a set of sample nodes from the RDF dataset and outputs a so called uniform shape constraint that uses only a restricted number of features of ShEx and SHACL. The uniform constraint can either be satisfied by all the nodes in the sample, in which case it is the unique most specific such constraint. Or it can take into account possible noise in the data and be satisfied by most sample nodes except for a small number of them consistent with a predefined error threshold. As a next contribution, we introduce schema patterns as a general mechanism for guiding and tuning our schema construction algorithm. A schema pattern can be provided as input to the algorithm, in which case the latter can output a complex shapes schema with nested constraints and references between shapes. In particular, we show how to construct a schema pattern from an existing ontology so that the automatically constructed schema contains one shape for each class in the ontology and integrates the predicates' domain and range axioms. We finally describe an interactive workflow that allows to construct an arbitrarily complex schema with the help of appropriate schema edition operators, statistical information about the data, a comprehensive visualization of validation results, and integrated calls to the algorithm for automatic schema construction.

Most of the ideas presented here are already implemented in the open source java tool *ShapeDesigner* available at <https://gitlab.inria.fr/jdusart/shexjapp>.

**Related work** RDF2Graph [2] performs structural graph summarization on knowledge bases that can be exported to RDF, OWL, XGML or ShEx. By default, this tool associates shapes to classes, taking into account class hierarchies. However, some tuning is possible, allowing to define constraints for sets of nodes that do not belong to the same class. It is limited w.r.t. expressible cardinalities but considers incoming and outgoing links in the neighbourhood of nodes. The resulting summary can be exported to ShEx or visualized as a graphical network, but none of these options are intended to be edited or refined by the user.

RDF2Graph has been used as a base to perform shape construction in Wikidata [16], using a publicly available tool<sup>4</sup>. This tool allows to specify an initial custom set of focus nodes to be associated with a shape. It can relate other shapes to the initial one, being those shapes constructed w.r.t nodes of the same class. It ignores triples in reified models, thus the schema of the metadata associated to a given piece of knowledge in Wikidata is missed.

In [4] and [7] the same authors propose two tools for automatic generation of ShEx schemas, which can be partially translated to SHACL in [7]. The tool in [4] takes as input a set of classes of a Knowledge Base (KB) and produces a ShEx shape for each one computing the immediate neighbourhood of the instances of that class. It uses statistical methods to choose the neighbourhood constraints, being able to discard the infrequent constraints and providing statistical annotations about them. The approach presented in [7] is independent of classes and the resulting constraints go beyond the immediate neighbourhood, but it is more limited w.r.t. cardinalities and less parametrizable. None of those proposals define an interaction process, and they are both less parametrizable than our method w.r.t. the structure of the schemas to be constructed.

The tool presented in [12] detects sets of nodes of an RDF graph that match some patterns, then combines those patterns to construct more complex OWL 2 EL class expressions and axioms. The latter can be exported to SHACL. The node groupings used for defining a shape are not necessarily related to the same class nor selected by the user, but they are automatically detected by the mining algorithm.

ABSTAT [13] is an ontology-driven tool able to produce semantic profiles of a given Knowledge Base (KB). Those profiles gather statistics about type and cardinality of features in the immediate incoming or outgoing neighbourhood of a set of target nodes. In [15], ABSTAT's profiles are translated to SHACL shapes associated to the instances of some target class.

PaTyBRED [11] is a hybrid classifier method which relies on path and type relations for the detection of errors in KBs. In [10], the authors train decision trees using PaTyBRED to generate SHACL constraints. The core idea is to detect examples qualified as highly erroneous, turn them into logical conditions and use their negation as a constraint.

The novelties of our approach compared to those cited above are (1) we include a strategy to integrate the user in the refinement of the automatically constructed schemas; and (2) we introduce schema patterns as an expressive parametrization mechanism. This allows to feed the schema construction algorithm with useful information coming from an ontology, or with expert knowledge.

*Organisation* In Sect. 2 we introduce a shapes schema language as an abstraction of ShEx and SHACL. In Sect. 3 we define uniform schemas as restriction of the language and explain the construction of the most specific and the largely accepted consensus uniform constraints. Schema patterns are defined in Sect. 4

---

<sup>4</sup><https://tools.wmflabs.org/wd-shex-infer/> retrieved on 2019/04/11.

and the interactive workflow in Sect. 5. We also discuss there the current status of the implementation of the method presented in this paper. We finally conclude and discuss future work in Sect. 6. This submission is accompanied by a long version extended with two appendices.

## 2 The Shape Constraint Language

We assume that the reader is familiar with RDF and has at least some basic understanding of the SHACL and ShEx constraint languages. We start by fixing notations related to RDF graphs then present the shapes constraint language.

Denote by **IRI** the set of IRIs, by **Lit** the set of literal values, and by **Blank** the set of blank nodes. Recall that an RDF graph **G** is a set of triples in  $(\text{IRI} \cup \text{Blank}) \times \text{IRI} \times (\text{IRI} \cup \text{Lit} \cup \text{Blank})$ . If  $(s, p, o)$  is a triple, then  $s$  is the subject,  $p$  the predicate, and  $o$  the object of the triple. A graph **G** will be clear from the context all along the paper. The elements of  $\text{IRI} \cup \text{Lit} \cup \text{Blank}$  that appear in subject or object position of some triple of **G** are called *nodes*, and the elements that appear in predicate position are called *predicates* or also *properties*. If  $N$  is a set of nodes, we define  $\text{props}(N) = \{p \mid \exists (s, p, o) \in \mathbf{G} \text{ s.t. } s \in N\}$ , i.e. the set of properties of the nodes in  $N$ . The *neighbourhood* of a node  $n$  is the set of triples of **G** with subject  $n$ . Node  $n'$  is called a *p-neighbour* of node  $n$  if  $(n, p, n')$  is a triple in **G**. Throughout the paper we use IRI prefixes without defining them and assume that different prefixes (e.g. `p:` and `q:`) always have different definitions.

SHACL [6] and ShEx [14] are the two main languages used for defining constraints for RDF graphs. See [9] for an introduction to these languages. They both use the notion of *shape* in order to describe or constraint the neighbourhood of nodes in RDF graphs, as illustrated in Example 1.

*Example 1 (The shapes `<Person>` and `<Date>`).* The shape `<Person>` requires 1) an `rdf:type` link to `foaf:Person`; 2) zero or more `owl:sameAs` links to some IRI; 3) a `foaf:name` link and a `foaf:familyName` link to string literals; and 4) a `bio:birth` link to a `xsd:gYear` or a `rdgr2:dateOfBirth` link to a node that satisfies shape `<Date>`, but not both. On the other hand, shape `<Date>` requires 1) an `rdf:type` link to `time:Instant`; and 2) an `rdfs:label` that is an integer value. This is how the constraint for shape `<Person>` would be written in the shapes constraint language.

```
{  rdf:type [foaf:Person] {1;1};      owl:sameAs IRI {0;*} ;
   foaf:name xsd:string {1;1};        foaf:familyName xsd:string {1;1} ;
   ( bio:birth xsd:gYear {1;1} | rdgr2:dateOfBirth @<Date> {1;1} ) }
```

In Appendix A we show how these shapes are expressed in ShEx and SHACL.

In Fig. 1 we provide some turtle content with a couple of nodes to be validated against the shape `<Person>`. `:virginia` conforms with `<Person>`, since it includes all the requested features regarding outgoing links. Shapes are *open* in the sense that they allow nodes to have outgoing links with predicates that are

not mentioned in the shape, as for instance the property `rdgr2:placeOfBirth` of `:virginia. :william` does not conform, since the information about Williams' name is provided using `schema:name` instead of `foaf:name` and `foaf:familyName`. In addition, `:william` also violates the requirement that only one among `bio:birth` or `rdgr2:dateOfBirth` links should be present.  $\square$

```

ex:virginia a foaf:Person ;      ex:william a foaf:Person ;
  foaf:name "Virginia" ;          owl:sameAs wd:Q692 ;
  foaf:familyName "Woolf" ;       schema:name "William Shakespeare" ;
  bio:birth "1882"^^xsd:gYear ;   rdgr2:dateOfBirth bnf:1564 ;
  rdgr2:placeOfBirth "London" .   bio:birth "1564"^^xsd:gYear .

bnf:1564 a time:Instant ; rdfs:label 1564 .

```

**Fig. 1.** Nodes of type `foaf:Person`

We now introduce the abstract Schapes Constraint Language (SCL) that containing the main features of ShEx and SHACL and has a quite straightforward translation to both of them described in Appendix A of the long version. Its syntax is very similar to ShEx compact syntax. A *shape constraint* `Constr` is defined by the abstract syntax in Fig. 2 using Extended Backus–Naur Form (EBNF) where the symbol `*` denotes the repetition of an element zero or more times. We next explain the semantics of SCL.

<code>Constr</code>	<code>::= ValueConstraints '{' NeighbConstr '}'</code>
<code>ValueConstraints</code>	<code>::= ValueConstr ( 'AND' ValueConstr )*   <math>\varepsilon</math></code>
<code>ValueConstr</code>	<code>::= XSD   'lit'   'nonlit'   'blank'   'iri'   Prefix   '[' ListValues ']'   'any'</code>
<code>XSD</code>	<code>::= an XML Schema type constant</code>
<code>Prefix</code>	<code>::= an IRI interpreted as a prefix or an RDF prefix with its terminating :</code>
<code>ListValues</code>	<code>::= a list of comma separated IRIs or literals</code>
<code>NeighbConstr</code>	<code>::= Conjunct ( ';' Conjunct )*</code>
<code>Conjunct</code>	<code>::= '(' Choice ')'   TripleConstr</code>
<code>Choice</code>	<code>::= TripleConstr ( ' ' TripleConstr )*</code>
<code>TripleConstr</code>	<code>::= Pred ObjectConstr '{' <i>min</i> ';' <i>max</i> '}'</code>
<code>ObjectConstr</code>	<code>::= '@' ShapeLabel   ValueConstr   '{' NeighbConstr '}'</code>
<code>Pred</code>	<code>::= an IRI</code>
<code>ShapeLabel</code>	<code>::= a label in <b>L</b></code>

**Fig. 2.** Syntax of the constraint language.

A `Constr` is satisfied by a node if this node and its neighbourhood satisfy both the value and the neighbourhood constraints. Each `ValueConstr` restricts the

value of a node to be one of the options in **ValueConstr**, where *lit* denotes a literal, *nonlit* a non-literal, *blank* a blank node, *iri* any IRI, and *any* imposes no value restriction, **XSD** defines a required XML Schema Definition type of the value, **Prefix** imposes it to be an IRI with the given prefix, and **ListValues** enumerates the admissible values. A **NeighbConstr** is satisfied by a node if its neighbourhood satisfies all the conjuncts. A **Choice** is satisfied when just one of the **TripleConstrs** composing it is satisfied. A **TripleConstr** is to be satisfied by the triples in the neighbourhood of the node having the specified predicate. The number of such triples must fit within the cardinality bounds given by *min* and *max*, where *min* is a natural and *max* is either a natural or the special value '\*' for *unbounded* cardinality. The objects of those triples should be nodes satisfying the specified **ObjectConstr**, which can be a simple **ValueConstr** or another shape. The latter is either a reference to **ShapeLabel** or a nested shape with {**NeighbConstr**}. Note finally that a **NeighbConstr** constrains only the triples in the neighbourhood which predicate is mentioned in one of its **TripleConstr**. All other predicates can be present without restriction on their object value or cardinality.

A *schema* is a pair  $\mathbf{S} = (\mathbf{L}, \mathbf{def})$  where  $\mathbf{L} \subseteq \text{IRI}$  is a finite set of shape labels and **def** associates with every label in **L** its definition which is a shape constraint. How an RDF graph should conform to a schema is defined by a validation target. A *validation target* for schema  $\mathbf{S} = (\mathbf{L}, \mathbf{def})$  and graph **G** is a pair  $\mathbf{T} = (\mathbf{L}, \mathbf{tgt})$  where **tgt** associates a set of nodes of **G** with every shape label from **L**. We say that graph **G** *satisfies* schema **S** with target **T** if for any shape label *L* in **L**, all nodes in **tgt**(*L*) satisfy the constraint **def**(*L*). Most typically, a target is specified using a class from an ontology or a SPARQL query. For instance, one might want to validate all nodes with class *City* and population greater than 1 million people against a shape **<BigCity>**. In this paper we consider non-recursive schemas only that do not allow the definition of a shape label to refer to itself neither directly nor transitively.

### 3 Automatic Construction of Uniform Constraints

Consider a fixed RDF graph **G**. In this section we address the problem of automatically constructing a shape constraint to be satisfied by a given sample set of nodes of **G**. We define uniform constraints which enjoy the nice property that for any set of sample nodes *N*, there exists a unique most specific uniform constraint satisfied by all the nodes in *N* (see Sect. 3.1). We also explain how to construct a uniform constraint satisfied by most but not necessarily all sample nodes, thus allowing to take into account noise in the sample data (see Sect. 3.2).

A *uniform constraint* is defined by the syntax in Fig. 3. We additionally require that the **UTripleConstrs** that appear in the same **UNeighbConstr** are without repeated properties, i.e. their predicates are pairwise distinct. Intuitively, uniform constraints allow to enumerate the predicates to appear in the neighbourhood of a node and to specify for each predicate a constraint for its value and its cardinality among a restricted choice of four possible cardinalities. Re-

```

UConstr      ::= '{' UNeighbConstr '}'
UNeighbConstr ::= UTripConstr ( ';' UTripConstr )*
UTripConstr  ::= Pred UValueConstr UCard
UValueConstr ::= XSD | 'lit' | 'nonlit' | 'blank' | 'iri' | Prefix | '[' UValue ']' | 'any'
UValue       ::= single IRI or literal
UCard        ::= '{0; *}' | '{0; 1}' | '{1; 1}' | '{1; *}'

```

**Fig. 3.** Syntax of uniform constraints.

mark that uniform constraints belong to the SCL language defined in Sect. 2.

### 3.1 Most Specific Constraint

Let  $F$  be the set of all constraints definable by `ValueConstrs` and let  $\preceq$  be the (partial) ordering relation over  $F$  defined by  $V_1 \preceq V_2$  if all nodes that satisfy  $V_1$  also satisfy  $V_2$ . Then for any prefix  $pr$ : and for any XSD datatype  $X$  it holds

$$pr: \preceq \text{iri} \preceq \text{nonlit} \preceq \text{any} \quad \text{blank} \preceq \text{nonlit} \quad X \preceq \text{lit} \preceq \text{any}$$

Moreover,  $\preceq$  is defined for prefix constraints by  $p: \preceq q$ : if  $q$  is a prefix of  $p$ , and also for XSD datatypes, for instance `int`  $\preceq$  `decimal`. Finally,  $\preceq$  is defined for `UValue` constraints by  $[r] \preceq X$  if  $r$  is a literal and  $X$  is its XSD type;  $[r] \preceq p$ : if  $r$  is an IRI and  $p$ : is its longest prefix, and  $[r] \preceq \text{blank}$  if  $r$  is a blank node.

Let  $K$  be the set of uniform cardinalities, then the subset relation  $\subseteq$  on intervals defines an ordering relation over  $K$ . It is not hard to show that

**Lemma 1.**  $(F, \preceq)$  is an upper semilattice with least upper bound  $\bigvee F = \text{any}$ .  $(K, \subseteq)$  is a complete lattice with least upper bound  $\bigvee K = \{0; *\}$  and greatest lower bound  $\bigwedge K = \{1; 1\}$ .

Therefore, for any set of RDF nodes  $R$ , there exists a unique most specific uniform value constraint satisfied by all nodes in  $R$  that we denote by  $\bigvee R$ . Similarly, for any set  $U$  of natural numbers, there exists a unique most specific uniform cardinality that includes  $U$  and we denote it  $\bigvee U$ .

**Definition 1 (Most specific constraint).** Let  $N \subseteq \text{Nodes}(G)$  be a sample set of nodes. The most specific constraint associated with  $N$  is the uniform constraint  $\text{msc}(N)$  that contains exactly one `TripleConstr`  $p \vee C$  for every predicate  $p \in \text{props}(N)$  whose value constraint  $V$  and cardinality  $C$  satisfy:

- $V = \bigvee N_p$  where  $N_p$  is the set of  $p$ -neighbours of  $N$ ;
- $C = \bigvee U$  where  $U = \{|(n, p, n') \in \mathbf{G}| \mid n \in N\}$ .

It easily follows from the definition that

**Lemma 2.** For any sample  $N \subseteq \text{Nodes}(\mathbf{G})$ , it holds that  $\mathbf{G}$  satisfies the schema  $(\{S\}, [S \mapsto \text{msc}(N)])$  with target  $(\{S\}, [S \mapsto N])$ .

The schema defined in Lemma 2 is the most precise schema using uniform constraints that one can construct for matching with the nodes of the sample  $N$ .



### 3.2 Largely Accepted Consensus Constraint

The largely accepted consensus constraint allows to account for potential errors in the neighbourhoods of the sample nodes. Imagine that most  $p$ -neighbours of the nodes in  $N$  have an IRI value, except for a very small number of them whose value is a literal. In this case,  $msc(N)$  will consider **any** as value constraint for  $p$ , while it is also possible that the literal  $p$ -neighbours are errors in the data.

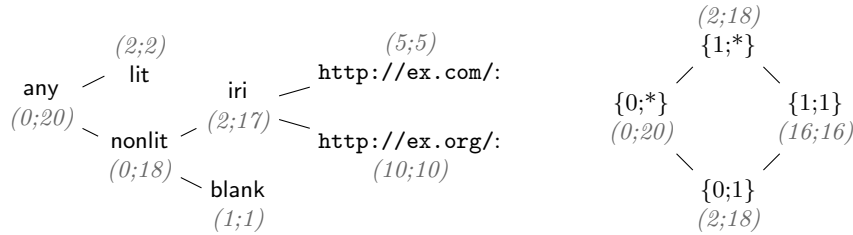
We start by presenting a decision making algorithm used in the definition of the largely accepted consensus constraint. Assume a non-empty set  $O$  of options to choose from by a finite set of voters  $W$ , and let  $(O, \preceq)$  be an upper semilattice with least upper bound  $o_\top$ , where  $o \preceq o'$  indicates that if option  $o$  is acceptable for given voter, then  $o'$  is acceptable for him/her as well. Every individual  $n$  votes for its preferred option  $\pi(n) \in O$ , and let  $acc_\pi(o)$  be the number of voters that consider option  $o$  as acceptable. Formally, for any total function  $\pi : W \rightarrow O$ , we define  $acc_\pi(o) = \sum_{o' \preceq o} |\{n \mid \pi(n) = o'\}|$ . It is then easy to see that  $acc_\pi(o_\top)$  is the total number of votes, i.e.  $acc_\pi(o_\top) = |W|$ .

Let  $0.5 < t \leq 1$  be a threshold value. We define the consensus option that is acceptable for at least a proportion  $t$  of the voters. Let  $a = \min \left\{ acc_\pi(o) \mid \frac{acc_\pi(o)}{|W|} \geq t \right\}$  be the least value in the range of  $acc_\pi$  s.t.  $\frac{a}{|W|}$  is above the threshold. Then

$$consensus(O, \preceq, W, \pi, t) = \bigvee \{o \mid acc_\pi(o) = a\}.$$

Note that, by definition,  $consensus(O, \preceq, W, \pi, t)$  is uniquely defined for all semilattices  $(O, \preceq)$ , sets of voters  $W$ , vote functions  $\pi : W \rightarrow O$  and thresholds  $t$ .

*Example 2 (Consensus).* Below on the left, consider the tree (thus also upper semilattice)  $(F', \preceq)$  with  $F' \subseteq F$ , where  $V$  is a child of  $V'$  whenever  $V \preceq V'$ . Assume a set  $W$  of 20 voters and a voting function  $\pi : W \rightarrow F'$ . For every option  $V$  we give between parentheses  $(x; y)$  where  $x$  is the number of direct votes (i.e.  $x = |\{w \in W \mid \pi(w) = V\}|$ ) and  $y = acc_\pi(V)$ . Then for different values of the threshold we obtain:  $consensus(F', \preceq, W, \pi, \frac{17}{20}) = \text{iri}$ ,  $consensus(F', \preceq, W, \pi, \frac{18}{20}) = \text{nonlit}$  and  $consensus(F', \preceq, W, \pi, \frac{19}{20}) = \text{any}$ . Below on the right is presented the lattice  $K$ , again annotated with  $(x, y)$  as above for a set  $W'$  of 20 voters and a voting function  $\pi' : W' \rightarrow K$ . Then  $consensus(K, \subseteq, W', \pi', \frac{17}{20}) = \bigvee \{\{0; 1\}, \{1; *\}\} = \{0; *\}$ .



□

The largely accepted consensus constraint is defined using a consensus where the voters are (some of) the nodes of the sample  $N$ . For every  $p \in props(N)$ ,

let  $N_p \subseteq N$  be the set of sample nodes that have at least one  $p$ -neighbour and assume we are given:

- $\pi_{F,p} : N_p \rightarrow F$  a vote function that gives the preferred value constraint for property  $p$  of every sample node that has a  $p$ -neighbour;
- $\pi_{K,p} : N \rightarrow K$  a vote function that gives the preferred cardinality for property  $p$  of every sample node.

**Definition 2 (Largely accepted consensus constraint).** *Let  $N \subseteq \text{Nodes}(G)$  be a sample set of nodes and  $0 \leq e < 0.5$  be an error rate value. The largely accepted consensus constraint associated with  $N$  and assuming error rate  $e$  is the uniform constraint  $\text{lac}_e(N)$  that contains exactly one `TripleConstr`  $p \ V \ C$  for every predicate  $p \in \text{props}(N)$  and it has value constraint  $V = \text{consensus}(F, \preceq, N_p, \pi_{F,p}, 1 - e)$ , and cardinality  $C = \text{consensus}(K, \subseteq, N, \pi_{K,p}, 1 - e)$ .  $\square$*

The error rate value  $e$  is intuitively a bound on the proportion of sample nodes that could have expressed an incorrect<sup>5</sup> vote due to noise in the data. It is not hard to see that  $\text{lac}_0(N) = \text{msc}(N)$ .

Finally we define the vote functions  $\pi_{F,p}$  and  $\pi_{K,p}$ . For any node  $n \in N$  and any  $p \in \text{props}(N)$ ,  $\pi_{K,p}(n)$  is defined as follows, where  $m$  is the number of  $p$ -neighbours of  $n$ :

$$\pi_{K,p}(n) = \begin{cases} \{0; 1\} & \text{if } m = 0 \\ \{1; 1\} & \text{if } m = 1 \\ \{1; *\} & \text{if } m > 1 \end{cases}$$

As for  $\pi_{F,p}$ , it is itself defined as a consensus:  $\pi_{F,p}(n) = \text{consensus}(F, \preceq, M, \text{best-match}, 1 - e')$  where  $0 \leq e' < 0.5$  is an error rate that might be different from  $e$ , the voters  $M = \{n' \mid (n, p, n') \in \mathbf{G}\}$  are all  $p$ -neighbours of  $n$ , and the voting function  $\text{best-match} : M \rightarrow F$  is defined by  $\text{best-match}(n') = [n']$  (i.e. the singleton `ListValues` containing  $n'$ ).

*Discussion* The  $\text{msc}$  and  $\text{lac}$  constraints are implicitly parametrized by  $F$  that could be replaced by any upper semilattice. For instance,  $F = \{\text{any}, \text{blank}, \text{lit}\}$  might be sufficient for some use cases. We choose to present this particular value for  $F$  because it is useful in practice.

## 4 Schema patterns

The main limitation of uniform constraints is that they do not allow for nested neighbourhood constraints or references to other shapes. A possible alternative could be to construct a schema up to some fixed nesting depth given as parameter [16,7] but the same nesting depth is seldom appropriate to all shapes. We propose to use schema patterns instead. They allow for a fine tuning of automatic schema

<sup>5</sup>Incorrect w.r.t. a supposedly existing correct uniform constraint that the user aims at.

```

wd:Q37100 a wikibase:Item ;
  wdt:P17 wd:Q664 ;           # (ds) country: New Zealand
  p:P17 [ a wikibase:Statement ; # (qualified statement)
    ps:P17 wd:Q664 ;         # country: New Zealand
    prov:wasDerivedFrom [ pr:P248 wd:Q265049 ] # provenance
  ] ;
  wdt:P6 wd:Q597909 ;         # (ds) head of gov.: Phil Goff
  p:P6 [ a wikibase:Statement ; # (qualified statement)
    ps:P6 wd:Q597909 ;       # head of gov.: Phil Goff
    pq:P580 "2016-10-09T00:00:00Z"^^xsd:dateTime # start time
  ] ;
  p:P6 [ a wikibase:Statement ; # (qualified statement)
    ps:P6 wd:Q4116955 ;      # head of gov.: Len Brown
    pq:P580 "2010-01-01T00:00:00Z"^^xsd:dateTime ; # start time
    pq:P582 "2016-10-08T00:00:00Z"^^xsd:dateTime ; # end time
    prov:wasDerivedFrom [ pr:P143 wd:Q328 ] # provenance
  ] .

```

Fig. 4. Extract of the Wikidata entry for Auckland

construction by describing a general form for the target schema, but can also be used simply to restrict the namespaces of predicates of interest, or to give a particular role to some values such as the value of `rdf:type`.

We start by a motivating example based on Wikidata that we use for an informal description of schema patterns in Sect. 4.1. Then in Sec 4.2 we give the syntax of schema patterns.

#### 4.1 Informal Definition

Fig. 4 presents an extract of the Wikidata entry for the city of Auckland<sup>67</sup>. Entities in Wikidata are encoded by identifiers that start with Q, and properties by identifiers that start with P. The entry in Fig. 4 includes a statements for the property country (P17) with value New Zealand (Q664), and two statements for the property head of government (P6) with values Phil Goff (Q597909) and Len Brown (Q4116955). The most significant value of each property is given as a direct statements (ds) (prefix `wdt:`), while richer information is presented in a qualified `wikibase:Statement` in which the claim is accompanied by context or scope information (under which conditions the claim holds), as well as meta information such as provenance. Note that the structures of such statements are similar. They usually contain property statements (`ps:`), qualifiers (`pq:`) and provenance information.

In Fig. 5 is presented the  $msc(N)$  constraint where the sample  $N$  contains the Wikidata entries for Auckland, Monterey, Vienna and Kobe<sup>7</sup>, which is clearly too

<sup>6</sup>From <https://www.wikidata.org/entity/Q37100.ttl> retrieved on 2019/03/27.

<sup>7</sup>restricted to the properties `rdf:type`, P17 and P6.

```

<City> {
  a [ wikibase:Item ] {1;1} ;
  wdt:P17 wd: {1;1} ;      p:P17 blank {1;*} ;
  wdt:P6  wd: {0;1} ;      p:P6  blank {0;*} }

```

Fig. 5. Most specific constraint for a sample of Wikidata cities

```

<City> {
  a      wikibase:      {1;1} ;
  wdt:P17 wd:           {1;1} ; # exactly one direct country statement
  p:P17  @<Y_P17>      {1;*} ; # several country statements
  wdt:P6  wd:           {0;1} ; # optional direct h. gov. statement
  p:P6    @<Y_P6>      {0;*} } # 0 or more h. gov. statements

  <Y_P17> {
    a [ wikibase:Statement ] {1;1} ;
    ps:P17 wd:           {1;*} } # at least one country statement

  <Y_P6> {
    a [ wikibase:Statement ] {1;1} ;
    ps:P6  wd:           {1;1} ;
    pq:P580 xsd:dateTime {1;1} ; # mandatory start time
    pq:P582 xsd:dateTime {0;1} } # optional end time

```

Fig. 6. Schema  $S_{\text{city}}$  for Wikidata cities

poor to represent the actual structure of the data. A more appropriate schema  $S_{\text{city}}$  for cities in Wikidata<sup>7</sup> is given in Fig. 6. It accounts for the possible qualifiers of the different properties.

A schema similar to  $S_{\text{city}}$  can be automatically constructed using the schema pattern  $P_{\text{city}}$  in Fig. 7 as input of the constraint construction algorithm. A schema pattern is similar to a schema but it omits cardinalities, it allows predicate filters ( $p:$ ,  $\text{iri}$ ,  $\text{ps:}$ , ...) instead of predicates, it replaces value constraints by placeholders  $\_$  or  $[-]$ , and it replaces some shape names by variables ( $Y$ ).

The pattern  $P_{\text{city}}$  defines one non-variable shape label  $\langle\text{City}\rangle$ , therefore it will be used with one set of sample nodes, say  $N_{\text{city}}$ , in order to construct a shapes schema. Denote  $P_{\text{city}}$  the pattern for  $\langle\text{City}\rangle$  and assume that  $N_{\text{city}}$  contains the single node  $\text{wd:Q37100}$  from Fig. 4. We start by matching all properties  $q$  in  $\text{props}(N_{\text{city}})$  with the predicate filters that appear in the  $P_{\text{city}}$ , and every such property  $q$  is matched with the most specific possible filter. That is  $p:P17$  and  $p:P6$  are matched with  $p:$  and we write  $\text{matched}(p:) = \{p:P16, p:6\}$ , while  $\text{matched}(\text{iri}) = \{\text{wdt:P17}, \text{wdt:P6}, \text{rdf:type}\}$ . Now every property in  $\text{matched}(p:)$  yields a triple constraint, and the pattern requires that its object is  $\text{@Y}$ , a shape reference to a freshly created shape label. We can see that schema  $S_{\text{city}}$  in Fig. 6 contains indeed two triple constraints  $p:P17 \text{ @<Y_P17>}$  and  $p:P6 \text{ @<Y_P6>}$ . The cardinalities of these triple constraints are not relevant here as they were obtained by considering a bigger sample set  $N_{\text{city}}$ . In the

general case, any property  $q \in \text{matched}(p:)$  defines its corresponding sample  $N_q = \{n' \mid \exists n \in N_{\text{city}}. (n, q, n') \in \mathbf{G}\}$  composed of the  $q$ -neighbours of the nodes in  $N_{\text{city}}$ . This sample is used to determine the cardinality of the triple constraint on the one hand, and as a sample for the definition of shape label  $\langle Y.q \rangle$  on the other hand. Similarly, every property in  $\text{matched}(\text{iri})$  yields a triple constraint which object constraint is a value constraint, as indicated by the placeholder  $\_$ . Now regarding the pattern  $P_Y$  for  $Y$ , we already saw that it defines two different shapes as defined by  $\text{matched}(p:)$ , and each of these has its associated sample. Remark that  $P_Y$  does not contain filter  $\text{iri}$ , therefore some of the properties of the sample nodes might match none of the filters. These properties will not be considered. For instance, property `prop:wasDerivedFrom` appears in the sample nodes in Fig. 4, but not in the schema  $\mathbf{S}_{\text{city}}$ . We finally explain the triple constraint `rdf:type [_]` in  $P_Y$ . Its predicate holder is not a filter but a property, so it will be matched only by this same property. Its object constraint is the placeholder `[_]` indicating that we want to construct a `ListValues` constraint for the set of values. In the example this is the singleton list that contains `wikibase:Statement`.

```

<City> {          # Shape <City> contains
  p:  @Y ;        # Triple constraints to new shapes for properties p:
  iri  _ _ }      # Triple constraints with value constraints for all other properties
Y {              # The new shapes for properties p: contain
  rdf:type  [_] ; # Triple constraint with list of values for rdf:type
  ps:  _ _ ;     # Triple constraints with value constraints for all properties ps:
  pq:  _ _ ; }   # Triple constraints with value constraints for all properties pq:

```

Fig. 7. Schema pattern  $\mathbf{P}_{\text{city}}$  for Wikidata cities

As we saw, schema patterns are particularly fitted for the Wikidata dataset as it strongly relies on reification. But they are also useful for all datasets that use reification [5] or repetitive local structure. Simpler schema patterns can also be used for e.g. restraining the properties of interest.

## 4.2 Formal Definition

Assume a countable set  $\mathcal{V}_{\text{sch}}$  of shape label variables ranging over  $X, Y, \dots$ . A *constraint pattern* (or pattern for short) is defined by the syntax in Fig. 8. A *schema pattern* is a triple  $\mathbf{P} = (\mathbf{PL}, \mathbf{pdef}, \mathbf{sample})$  where:

- $\mathbf{PL} \subseteq \mathcal{L} \cup \mathcal{V}_{\text{sch}}$  is a finite set of shape labels and shape label variables;
- with every label or variable in  $\mathbf{PL}$ ,  $\mathbf{pdef}$  associates a shape pattern that uses shape variables from  $\mathbf{PL} \cap \mathcal{V}_{\text{sch}}$  only;
- $\mathbf{sample}$  associates a set of nodes from  $\mathbf{G}$  with every label in  $\mathbf{PL} \cap \mathcal{L}$ .

Pattern	::= '{' PNeighbConstr '}'
PNeighbConstr	::= PTripleConstr ( ';' PTripleConstr )*
PTripleConstr	::= PredHolder ObjectConstrHolder
PredHolder	::= PredFilter   Pred
PredFilter	::= Prefix   'iri'
ObjectConstrHolder	::= '@' SVar   '._'   '[_...]'   '{' PNeighbConstr '}'
SVar	::= a variable in $\mathcal{V}_{\text{sch}}$

**Fig. 8.** Syntax of constraint patterns.

A schema pattern defines a most specific uniform schema (or a largely accepted consensus schema) which definition is given in Appendix B of the long version due to space limitations.

### 4.3 Patterns for Ontologies

We show here how a schema pattern can be used to encode some of the information available in an existing ontology that might be relevant for the schema construction. Consider the toy RDFS schema ontology  $R$  that defines three classes `:Human`, `:Teacher` and `:Subject` with `:Teacher` subclass of `:Human`, and three properties `:name`, `:teaches`, `:description` with axioms

```

:name      rdfs:domain :Human.      :description rdfs:domain :Subject.
:teaches   rdfs:domain :Teacher.    :teaches      rdfs:range  :Subject.

```

We can automatically retrieve the following schema pattern

```

<Human>    { a [ _ _ ] ; :name _ _ }
<Teacher>  { a [ _ _ ] ; :name _ _ ; :teaches _ _ }
<Subject>  { a [ _ _ ] ; :description _ _ }

```

with `tgt(<Human>)` defined by the SPARQL query `SELECT ?x WHERE ?x a :Human.` and similarly for `<Teacher>` and `<Subject>` shapes. The shapes schema constructed with this pattern will contain only the properties of the ontology vocabulary attached to their respective domains.

## 5 Interactive Improvement of the Schema

We are convinced that an optimal shapes schema, i.e. a schema that encodes correctly the requirements for a particular application, cannot be obtained by a fully automatic process. Automatically constructed schemas are necessarily simple: if we allowed the choice operator or repeated properties<sup>8</sup> in uniform constraints, then a unique most specific constraint would not exist. Schema patterns (or other forms of parametrization) can bring a partial solution, but they

<sup>8</sup>that is, allow two triple constraints with same predicate in the same neighbourhood constraint

may require user expertise and good knowledge about the data, which is often the reason why we want to build a schema at the first place: a chicken-and-egg situation.

Our solution is to let the user build the schema but within the interactive workflow described hereafter that gives access to automatic schema construction and to useful information about the data, and provides appropriate feedback.

Assume an RDF graph  $\mathbf{G}$  for which we want to build a schema. The interactive workflow maintains a schema  $\mathbf{S} = (\mathbf{L}, \mathbf{def})$  together with associated validation target  $\mathbf{T} = (\mathbf{L}, \mathbf{tgt})$  and schema pattern  $\mathbf{P} = (\mathbf{PL}, \mathbf{pdef}, \mathbf{sample})$  with  $\mathbf{sample} = \mathbf{tgt}$ . A target  $\mathbf{tgt}(L)$  is given as a tuple  $(Q, N^+, N^-)$  of a SPARQL query  $Q$  to be executed against  $\mathbf{G}$  and sets of nodes  $N^+$  and  $N^-$ , to be added, resp. removed from the result of query  $Q$ . The schema and the targets can be modified using the following editing operations:

- modify  $\mathbf{def}(L)$  for shape label  $L$  using these actions:
  - add or remove a **Conjunct** or **ValueConstraints**;
  - group two or more triple constraints into a **Choice**, or split a choice;
  - split a **TripleConstr** of the form  $p \ V \ C$  in two **TripleConstrs**  $p \ V_1 \ C_1$  and  $p \ V_2 \ C_2$  having the same predicate, or regroup two such triple constraints;
  - change the cardinality or the **ObjectConstr** of some triple constraints with predicate  $p$ . In particular, the object constraint can be replaced by the automatically constructed  $\{msc(N')\}$  or  $\{lac(N')\}$ , where  $N'$  is the set of  $p$ -neighbours of  $\mathbf{tgt}(L)$ ;
- modify  $\mathbf{tgt}(L)$  for shape label  $L$ . This action is triggered automatically when some triple constraint is changed and becomes of the form  $p \ @L \ C$  (with shape reference to  $L$ ). Then the set  $N'$  of all IRI  $p$ -neighbours of  $\mathbf{tgt}(L)$  is added to  $\mathbf{tgt}(L')$ . In the latter case, the **NeighbConstr**  $M$  is automatically constructed as  $msc(N')$  or  $lac(N')$ .
- add a new shape label  $L$  to  $\mathbf{L}$  with corresponding target  $\mathbf{tgt}(L) = N$  and pattern  $\mathbf{pdef}(L)$ , which triggers the automatic construction of  $msc(N)$  or  $lac(N)$  using  $\mathbf{pdef}(L)$  as pattern and assigns the result to  $\mathbf{def}(L)$ .

These operations allow to construct an arbitrarily complex SCL schema starting e.g. from an automatically constructed one.

At any time the user can visualize the result of validating  $\mathbf{G}$  against  $\mathbf{S}$  with target  $\mathbf{T}$ . Additionally, the user has access to the following information:

- for any shape label  $L$  and triple constraint  $p \ V \ C$  that occurs in  $\mathbf{def}(L)$  or in  $\mathbf{pdef}(L)$ , and denoting  $N = \mathbf{tgt}(L)$ :
  - the number of nodes in  $N$  that have  $p$  as predicate;
  - the minimal, maximal and average number of occurrences of predicate  $p$  in the neighbourhood of  $N$ ;
  - the lattice of uniform value constraints  $(F, \preceq)$  annotated with  $acc_{\pi_F}(V')$  for every  $V' \in F$  (see Sect. 3.1);
- for any pair of predicates  $p, p'$ , the number of nodes from  $N$  in which  $p$  and  $p'$  co-occur; this is useful to detect possible **Choices**.

**Implementation** We have already implemented several components of our method, some of which are illustrated in the video attached to the submission:

- the algorithm for automatic construction of the most specific schema with only limited support of schema patterns tailored for Wikidata (video),
- simple visualisation of statistics about the data,
- enhanced visualisation of validation results that allows to navigate synchronously in the schema and in the graph to understand why a given node is valid or not w.r.t. a given constraint (video).

Currently our implementation is based on ShEx validation. In order to complete the implementation we need to: (1) complete the implementation of schema patterns; (2) add support for SHACL validation; (3) provide a nice visualisation of the information about the data; and (4) integrate the different components into a single application.

Fig. 9 shows a consolidated view of three of the already implemented components. We consider a sample with four Wikidata entries: Auckland, Monterey, Vienna and Kobe (restricted to a part of their predicates only). The top right panel contains the schema that was automatically constructed using a schema pattern similar to the one presented in Fig. 7. We changed a single cardinality in that schema that produced a validation error. In the bottom panel we see the list of validated nodes on which the error is highlighted. The top left panel shows part of the information provided to the user, i.e. the list of the properties of sample nodes and their frequencies of occurrence.

The screenshot displays a web-based tool interface for ShEx schema construction and validation. It is divided into several panels:

- Indications (Top Left):** A tree view showing data indications for properties p:P17 and p:P6, including their associated predicates and cardinalities.
- Schema Construction (Top Right):** A text area showing the automatically generated ShEx schema, with buttons for "Save ShEx" and "Validate ShEx".
- Validation Results (Bottom):** A table showing the results of validating sample nodes against the schema.

Node	Shape	Result	Message
wd:Q487315	<http://city.io/City>	false	[The number of matched triples do...
wd:Q37100	<http://city.io/City>	true	[]
wd:Q48320	<http://city.io/City>	true	[]
wd:Q1741	<http://city.io/City>	true	[]

**Fig. 9.** Screenshot of some parts of the tool



## 6 Conclusion and Future Work

We have presented a method that allows to define ShEx or SHACL schemas for existing datasets. The construction of simple (uniform) schemas is fully automatic and is accessible to novice users, while experts can supply their knowledge about the data using schema patterns as an expressive parametrization mechanism, and further edit the resulting schema. A prototype implementation of the method allows to extract and refine ShEx schemas for Wikidata. We are currently extending the implementation in order to integrate the still missing features: full support for schema patterns, translation to SHACL and integrated SHACL validation, a more readable representation of the statistics about the data using GUI widgets. As a future work we are planning to enhance schema patterns in order to capture richer information that comes from ontologies such as cardinalities, and also to allow to describe more complex references between shapes.

## References

1. Čebirić, Š., Goasdoué, F., Kondylakis, H., Kotzinos, D., Manolescu, I., Troullinou, G., Zneika, M.: Summarizing semantic graphs: a survey. *The VLDB Journal* (2018)
2. van Dam, Jesse, C., Koehorst, J.J., Schaap, P.J., Martins, V.A., Suarez-Diez, M.: RDF2Graph a tool to recover, understand and validate the ontology of an RDF resource. *Journal of Biomedical Semantics* **6** (2015)
3. De Meester, B., Heyvaert, P., Dimou, A., Verborgh, R.: Towards a uniform user interface for editing data shapes. In: 4th VOILA. vol. 2187, pp. 13–24 (2018)
4. Fernández-Álvarez, D., García-González, H., Frey, J., Hellmann, S., Labra Gayo, J.E.: Inference of Latent Shape Expressions Associated to DBpedia Ontology. In: *International Semantic Web Conference* (2018)
5. Frey, J., Müller, K., Hellmann, S., Rahm, E., Vidal, M.E.: Evaluation of Metadata Representations in RDF stores. *Semantic Web (Preprint)*, 1–25 (2017)
6. Knublauch, H., Ryman, A.: Shapes constraint language (SHACL). W3C Candidate Recommendation, W3C (July 2017), <https://www.w3.org/TR/shacl/>
7. Labra Gayo, J.E., Fernández-Álvarez, D., García-González, H.: RDFShape: An RDF playground based on Shapes. In: *Proceedings of ISWC* (2018)
8. Labra Gayo, J.E., García-González, H., Fernández-Álvarez, D., Prud’hommeaux, E.: Challenges in RDF Validation. In: Alor-Hernández, G., Sánchez-Cervantes, J.L., Rodríguez-González, A. (eds.) *Current Trends in Semantic Web Technologies: Theory and Practice*, chap. 6, pp. 121–151 (2018)
9. Labra Gayo, J.E., Prud’hommeaux, E., Boneva, I., Kontokostas, D.: *Validating RDF Data*, vol. 7. Morgan & Claypool Publishers LLC (2017)
10. Melo, A.: Automatic refinement of large-scale cross-domain knowledge graphs. Ph.D. thesis (2018)
11. Melo, A., Paulheim, H.: Detection of relation assertion errors in knowledge graphs. In: *Proceedings of the Knowledge Capture Conference*. p. 22. ACM (2017)
12. Potoniec, J., Jakubowski, P., Lawrynowicz, A.: Swift linked data miner: Mining OWL 2 EL class expressions directly from on-line rdf datasets. *Journal of Web Semantics First Look* (2017)

13. Principe, R.A.A., Spahiu, B., Palmonari, M., Rula, A., De Paoli, F., Maurino, A.: ABSTAT 1.0: Compute, Manage and Share Semantic Profiles of RDF Knowledge Graphs. In: European Semantic Web Conference. pp. 170–175 (2018)
14. Prud’hommeaux, E., Boneva, I., Labra Gayo, J.E., Gregg, K.: Shape Expressions Language (ShEx). W3C Shape Expressions Community Group Drat Report (2018)
15. Spahiu, B., Maurino, A., Palmonari, M.: Towards Improving the Quality of Knowledge Graphs with Data-driven Ontology Patterns and SHACL. In: WOP@ISWC (2018)
16. Werkmeister, L.: Schema Inference on Wikidata. Master Thesis (2018)

## A The Shape Constraint Language, SHACL and ShEx

We explain here how SCL is translated to SHACL and to ShEx, then we explain the small differences in semantics depending on the target language.

*Translation to SHACL* SCL’s `ValueConstrs` are represented using `sh:nodeKind` (for `lit`, `nonlit`, `blank`, `iri`), using `sh:datatype` for `XSD`, using `sh:pattern` for `Prefix`, and using `sh:in` for `ListValues`. A `NeighbConstr` yields a `sh:shape` that is a conjunction (`sh:and`) of the encodings of its `Conjuncts`. The `Choice` operator can be represented by `sh:or` or by `sh:xone` (which one to use is a parameter of the translation). `TripleConstr` is encoded using `sh:property` with `sh:path` constraint for its predicate, while its `ObjectConstr` is represented by `sh:node` if it is a shape reference or a nested `NeighbConstr`, and by the encoding of the `ValueConstr` otherwise. Cardinality is represented by `sh:minInclusive` and `sh:maxInclusive` as expected. Additionally, if two or more triple constraints in the same neighbourhood constraint have the same predicate, then all these triple constraints are encoded using `sh:qualifiedValueShape`, and their cardinalities using `sh:qualifiedMinCount` and `sh:qualifiedMaxCount`. Finally, SHACL shapes have associated target declarations that are not provided by the SCL schema but by its accompanying validation target. In Fig. 10 we give as example the shapes `<Person>` and `<Date>` from Sect. 2 written in SHACL.

*Translation to ShEx* The syntax of SCL is based on the compact syntax of ShEx, therefore the translation to ShEx is almost straightforward, except for `Conjunct` that becomes *EachOf* also denoted `;` in the compact syntax, and `Choice` that becomes *OneOf* denoted `|`. Additionally, cardinality `{1;1}` is omitted, cardinality `{0;*}` becomes `*`, `{1;*}` becomes `+` and `{0;1}` becomes `?`, while any other cardinality remains as it is. In Fig. 11 we represent the translation to ShEx of shapes `<Person>` and `<Date>` from Sect. 2.

*Discussion* The semantics of a SCL schema can be different if it is translated to ShEx or to SHACL. The two differences come from the choice operator and from the use of different triple constraints with the same property in the same neighbourhood constraint. This is not an issue for our approach because we make the hypothesis that the user constructs a schema for one of the two languages, not for both, depending on the application she or he has in mind and on her/his

```

<Person> a sh:NodeShape ;
  sh:property [ sh:path owl:sameAs ;   sh:nodeKind sh:IRI ;
               sh:minCount 0 ] ;
  sh:property [ sh:path foaf:name ;     sh:datatype xsd:string ;
               sh:minCount 1 ;   sh:minCount 1 ] ;
  sh:property [ sh:path foaf:familyName ; sh:datatype xsd:string ;
               sh:minCount 1 ;   sh:minCount 1 ] ;
  sh:property [ sh:path rdf:type;       sh:hasValue foaf:Person ;
               sh:minCount 1; sh:maxCount 1 ] ;
  sh:xone ( [ sh:property [ sh:path bio:birth ;   sh:datatype xsd:gYear ;
                        sh:minCount 1 ;   sh:maxCount 1 ] ]
            [ sh:property [ sh:path rdgr2:dateOfBirth ; sh:nodeKind sh:IRI ;
                        sh:minCount 1 ;   sh:maxCount 1 ; ] ] ) ;
  sh:closed false ;   sh:targetClass foaf:Person .

<Date> a sh:NodeShape ;
  sh:property [ sh:minCount 1 ;   sh:maxCount 1 ;
               sh:path rdf:type ; sh:hasValue time:Instant ] ;
  sh:property [ sh:minCount 1 ;   sh:maxCount 1 ;
               sh:path rdfs:label ; sh:datatype xsd:int ] ;
  sh:closed false;   sh:targetClass time:Instant .

```

Fig. 10. Shape &lt;Person&gt; in SHACL

```

<Person> { rdf:type [foaf:person] ;
          owl:sameAs IRI * ;
          foaf:name xsd:string ;
          foaf:familyName xsd:string ;
          ( bio:birth xsd:gYear | rdgr2:dateOfBirth @<Date> ) }
<Date> { rdf:type [time:Instant] ; rdfs:label xsd:int }

```

Fig. 11. Shape &lt;Person&gt; in ShEx

production suite. The validation results in the interactive tool are computed using the translation to the chosen target schema language, whichever it is. When the desired abstract SCL schema is ready, it is exported and can be used in production.

## B Formal Definition of Shape Patterns

A schema pattern is defined by the syntax in Fig. 8. We additionally require that (1) no **PredFilter** is used more than once in the same constraint pattern, and (2) every shape label variable  $Y \in \mathcal{V}_{\text{sch}}$  is used exactly once in the range of **pdef**.

Denote by  $E$  the set of **PredFilters** of constraint patterns. Remark that  $E$  is a subset of the set  $F$  of value constraints defined in Sect. 3, and that  $(E, \preceq)$  is an upper semilattice with least upper bound  $\bigvee E = \text{iri}$ . For any **PNeighbConstr**  $P$ ,

denote  $E(P)$  the set of predicate filters that appear in the direct **PTripleConstrs** of  $P$  (that is, excluding those that appear in some nested **PNeighbConstr**). Because  $E(P) \subseteq E$  and  $(E, \preceq)$  is a lattice, it holds that for every property  $q$  either no element of  $E(P)$  matches  $q$  and we write  $\text{match}(P, q) = \perp$ , or there is unique most precise  $f$  in  $E(P)$  s.t.  $[q] \preceq f$  and we write  $\text{match}(P, q) = f$ .

For a property  $q$ , denote  $E(P)_{q \preceq}$  the set of filters in  $E(P)$  that match  $q$ , i.e.  $E(P)_{q \preceq} = \{f \in E(P) \mid [q] \preceq f\}$ . For instance, if  $P$  is **pdef**( $Y$ ) from Fig. 7 then  $ps: \in E(P)$  and  $ps: \in E(P)_{ps:foo \preceq}$ .

Fix a pattern schema  $\mathbf{P} = (\mathbf{PL}, \mathbf{pdef}, \mathbf{sample})$  for the sequel. Its associated most specific schema with uniform constraints  $\text{msc}(\mathbf{P})$  is defined below. We point out first that  $\text{msc}(\mathbf{P})$  is not a uniform schema as it allows nested neighbourhood constraints and shape references, but likewise uniform schemas, its neighbourhood constraints satisfy do not use choice or repeated properties. So let  $\text{msc}(pat) = (\mathbf{L}, \mathbf{def})$ , we explain now which are the shape labels in  $\mathbf{L}$  and which are their definitions.

For every shape label  $L \in \mathbf{PL}$ ,  $L$  is in  $\mathbf{L}$  and its definition  $P = \mathbf{pdef}(L)$  together with its sample  $N = \mathbf{sample}(L)$  define a **NeighbConstr**  $S = \text{msc}(P, N)$ .  $S$  is a conjunct of **TripleConstrs** (without **Choices**). Denote by  $\text{filters}(P)$ , resp. by  $\text{predicates}(P)$ , the set of **Predicates**, resp. of **PredFilters**, that occur in the **PTripleConstrs** of  $P$ . For instance,  $\text{filters}(\mathbf{pdef}(Y)) = \{ps:, pq:\}$  and  $\text{predicates}(\mathbf{pdef}(Y)) = \{rdf:type\}$  for the pattern for shape  $Y$  in Fig. 7. Then let  $\text{props}(P, N)$  be the set that contains exactly the properties that appear in  $P$  and all properties that appear in the neighbourhood of  $N$  and match some of the filters in  $P$ .

$$\text{props}(P, N) = \text{predicates}(P) \cup \{q \in \text{props}(N) \mid \text{match}(P, q) \neq \perp\}.$$

Finally, for every predicate  $q \in \text{props}(P, N)$ , its associated sample  $s(q)$  is the set of  $q$ -neighbors of  $N$ .

Each  $q \in \text{props}(P, N)$  yields a triple constraint  $q \text{ } O \text{ } C$  which object constraint  $O$  and cardinality constraint  $C$  are defined as follows. Let  $H_q$  be the object constraint holder for  $q$  by  $P$ . That is, if  $q \in \text{predicates}(P)$ , then  $H_q$  is the object constraint holder associated with  $q$  in  $P$ . Otherwise,  $H_q$  is the object constraint holder associated with  $\text{match}(P, q)$ , the most precise predicate filter that matches  $q$ .

Now, the object constraint  $O$  is, depending on the form of  $H_q$ , and denoting  $N_q$  the set of  $q$ -neighbours of  $N$ :

- If  $H_q = \_$  (a value constraint holder), then  $O = \bigvee N_q$ .
- If  $H_q = [\_]$  (a list of values holder), then  $O$  is the list that contains exactly the values in  $N_q$ .
- If  $H_q = \{M\}$  (a nested neighbourhood constraint pattern), then  $M = \text{msc}(P, N_q)$ .
- If  $H_q$  is of the form  $@X$  (shape reference), then we create a fresh shape label  $\langle X\_q \rangle$  and  $O = @\langle X\_q \rangle$ . We add  $\langle X\_q \rangle$  to  $\mathbf{L}$  and its definition  $\mathbf{def}(\langle X\_q \rangle) = \text{msc}(\mathbf{pdef}(X), N_q)$ .

Because the pattern  $\mathbf{P}$  is not recursive (requirement (2) above), this procedure terminates and constructs a shapes schema.

We finally point out that schema patterns might suffer of the problem of “shrinking samples”. Even if the size of **sample**( $L$ ) for some  $L \in \mathbf{PL}$  is sufficient, its nested neighbourhood constraints are constructed using samples of  $q$ -neighbours of the nodes in  $N$  for the appropriate predicate  $q$ . If predicate  $q$  occurs only a few times in the neighbourhood of  $N$ , then the corresponding sample can be very small. In our implementation we warn the user if such situation occurs, and allow her to add additional nodes to that sample without necessarily augmenting **sample**( $L$ ) (for efficiency and readability reasons).